Analise Exploratória de Dados em Linguagem R

# Carregando os pacotes

library(corrplot)

## corrplot 0.89 loaded

library(gridExtra)  
library(ggplot2)

# Criando um Dataframe para teste

df <- data.frame(Estado =c("SP", "BA", "AM"),  
 Plano = c("Bronze", "Ouro", "Prata"),  
 Valor = c(500, 300, 150),  
 Total = c(1000, 300, 300),  
 Dependentes = c(2,1,2),  
 Naturalidade = c("BR", "BR", "US"),  
 Desconto = c("Sim", "Não", "Não"))  
df

## Estado Plano Valor Total Dependentes Naturalidade Desconto  
## 1 SP Bronze 500 1000 2 BR Sim  
## 2 BA Ouro 300 300 1 BR Não  
## 3 AM Prata 150 300 2 US Não

str(df)

## 'data.frame': 3 obs. of 7 variables:  
## $ Estado : chr "SP" "BA" "AM"  
## $ Plano : chr "Bronze" "Ouro" "Prata"  
## $ Valor : num 500 300 150  
## $ Total : num 1000 300 300  
## $ Dependentes : num 2 1 2  
## $ Naturalidade: chr "BR" "BR" "US"  
## $ Desconto : chr "Sim" "Não" "Não"

Os dados contém 7 colunas e 3 linhas.

# Limpeza dos Dados

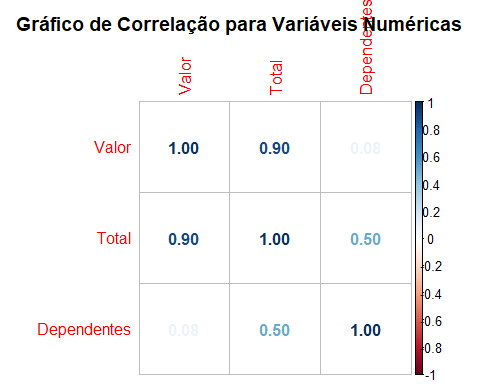
# Verificando valores ausentes

sapply(df, function(x) sum(is.na(x)))

## Estado Plano Valor Total Dependentes Naturalidade   
## 0 0 0 0 0 0   
## Desconto   
## 0

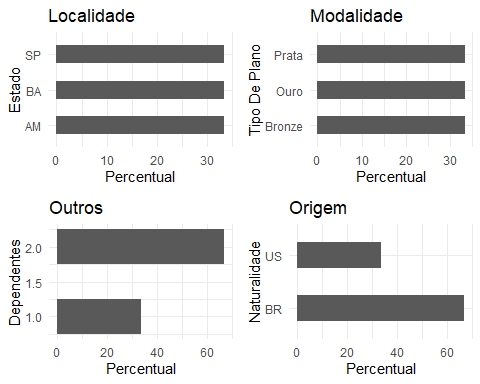
# Análise Preliminar dos Dados

numeric.var <- sapply(df, is.numeric)  
corr.matrix <- cor(df[,numeric.var])  
corrplot(corr.matrix, main="\n\nGráfico de Correlação para Variáveis Numéricas", method="number")



Correlação entre as variáveis numericas.

p1 <- ggplot(df, aes(x=Estado)) + ggtitle("Localidade") + xlab("Estado") +  
 geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentual") + coord\_flip() + theme\_minimal()  
p2 <- ggplot(df, aes(x=Plano)) + ggtitle("Modalidade") + xlab("Tipo De Plano") +   
 geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentual") + coord\_flip() + theme\_minimal()  
p3 <- ggplot(df, aes(x=Dependentes)) + ggtitle("Outros") + xlab("Dependentes") +   
 geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentual") + coord\_flip() + theme\_minimal()  
p4 <- ggplot(df, aes(x=Naturalidade)) + ggtitle("Origem") + xlab("Naturalidade") +   
 geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentual") + coord\_flip() + theme\_minimal()  
grid.arrange(p1, p2, p3, p4, ncol=2)



Distribuição das variáveis categóricas, todas elas serão mantidas para análise posterior.